Eléments de sciences ouvertes

Laurent Romary

Eléments de sciences ouvertes

Laurent Romary, Inria ALMAnaCH

Séminaire DIM MAP – 15 mai 2020

Introduction à la journée

Objectif

- Améliorer les pratiques de gestion des données en sciences du patrimoine
- Recueil d'expérience sur les pratiques des projets soutenus par le DIM MAP
  - Présentation des projets sous l'angle des données
    - Acteurs, sources, documentation, hébergement, ouverture, réutilisation
  - Bilan des pratiques stables et recommandations
  - Expression des attentes de chacun
    - Priorités d'action (DIM MAP, national/CoSO)
- Faire le point sur les principes et initiatives en matière de science ouverte
  - Présentation du cadre courant (légal, outils, gestion du workflow en sciences du patrimoine)
- Séance finale de mise à niveau d'égalité sous HAL

Ouvrir la science

Pourquoi ouvrir la science ?

- Fluidifier la circulation des résultats scientifiques
  - Croiser les informations, les valider, les reproduire
  - Contribuer à des débats sociétaux éclairés
- Vers une vision systémique/systématique de science ouverte
  - Ouvrir tout, autant que possible, et à toutes les étapes d'un processus de recherche
    - Propositions de projet
    - Données brutes, intermédiaires
    - Processus, modèles, formats (= méta-science)
    - Résultats, analyses, publications
- Pour autant, il n'y a pas une seule science ouverte
  - Disciplines, contraintes institutionnelles, possibilités techniques

In memoriam: John Tennant

Comment ouvrir la science ?

- Démarche individuelle et collective
  - Anticiper sur la dissémination des différentes composantes d'un projet scientifique
    - Intégration aux méthodes
  - La science ouverte comme partie intégrante des politiques d'établissements d'ESR
    - Lien avec les bilans d'activité, les évaluations, la communication institutionnelle
- Réfléchir à la nécessité d'une souveraineté numérique sous-jacente
  - L'ouverture repose sur l'existence d'hébergements numériques fiables et pérennes
  - Importance d'infrastructures publiques soutenant la science ouverte
- Projet politique
  - Intégrer la science ouverte dans la démarche des tutelles et financeurs de la recherche
    - Conditions de financement, plans de gestion des données
    - *Showing by doing*: ex. HCERES => https://hal-hceres.archives-ouvertes.fr
- Tenir compte du cadre légal

Que dit la loi ? - Principes généraux

- Ouverture des données par défaut
  - Données de la recherche comme document administratif
  - Responsabilité: institution, opérateur: chercheur
  - Cas spécifiques: diffusion obligatoire des données géographiques et environnementales
- Exceptions possibles
  - Propriété intellectuelle, données personnelles, considérations stratégiques ou de valorisation
  - Exceptions particulières pour les musées, archives, bibliothèques
- Cadre législatif
  - Loi n° 78-753 du 17 juillet 1978, CADA (Commission d'accès aux documents administratifs)
  - Directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 sur la réutilisation des informations du secteur public. PSI (*Public Sector Information*)
  - Directive européenne du 26 juin 2013 relative à la réutilisation des informations du secteur public
  - Loi dite Valter du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public
  - Loi pour une République numérique - 7 octobre 2016
  - Intégration dans le Code des relations entre le public et l'administration (livre III, titre II)

Code de la recherche - Article L533-4

I.-Lorsqu'un **écrit scientifique issu d'une activité de recherche financée au moins pour moitié** par des dotations de l'Etat, des collectivités territoriales ou des établissements publics, par des subventions d'agences de financement nationales ou par des fonds de l'Union européenne est publié dans un périodique paraissant au moins une fois par an, **son auteur dispose**, même après avoir accordé des droits exclusifs à un éditeur, **du droit de mettre à disposition gratuitement dans un format ouvert**, par voie numérique, sous réserve de l'accord des éventuels coauteurs, **la version finale de son manuscrit acceptée pour publication**, dès lors que l'éditeur met lui-même celle-ci gratuitement à disposition par voie numérique ou, à défaut, à l'expiration d'un délai courant à compter de la date de la première publication. Ce délai est au maximum de **six mois** pour une publication dans le domaine des sciences, de la technique et de la médecine et de **douze mois** dans celui des sciences humaines et sociales.

La version mise à disposition en application du premier alinéa ne peut faire l'objet d'une exploitation dans le cadre d'une activité d'édition à caractère commercial.

II.-Dès lors que **les données issues d'une activité de recherche** financée au moins pour moitié par des dotations de l'Etat, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont **pas protégées par un droit spécifique ou une réglementation particulière** et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, **leur réutilisation est libre**.

III.-L'éditeur d'un écrit scientifique mentionné au I ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication.

IV.-Les dispositions du présent article sont d'ordre public et toute clause contraire à celles-ci est réputée non écrite.

La plan national pour la science ouverte

- Annoncé par Frédérique Vidal, le 4 juillet 2018
- Contient trois volets:
  - Accès ouvert aux publications
    - Rôle prépondérant de HAL
  - Structurer et ouvrir les données de la recherche
    - « Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics »
  - Dynamique durable, européenne et internationale
    - écoles doctorales, opérateurs (ANR, HCERES, etc.)

L’expérience Inria

Participation

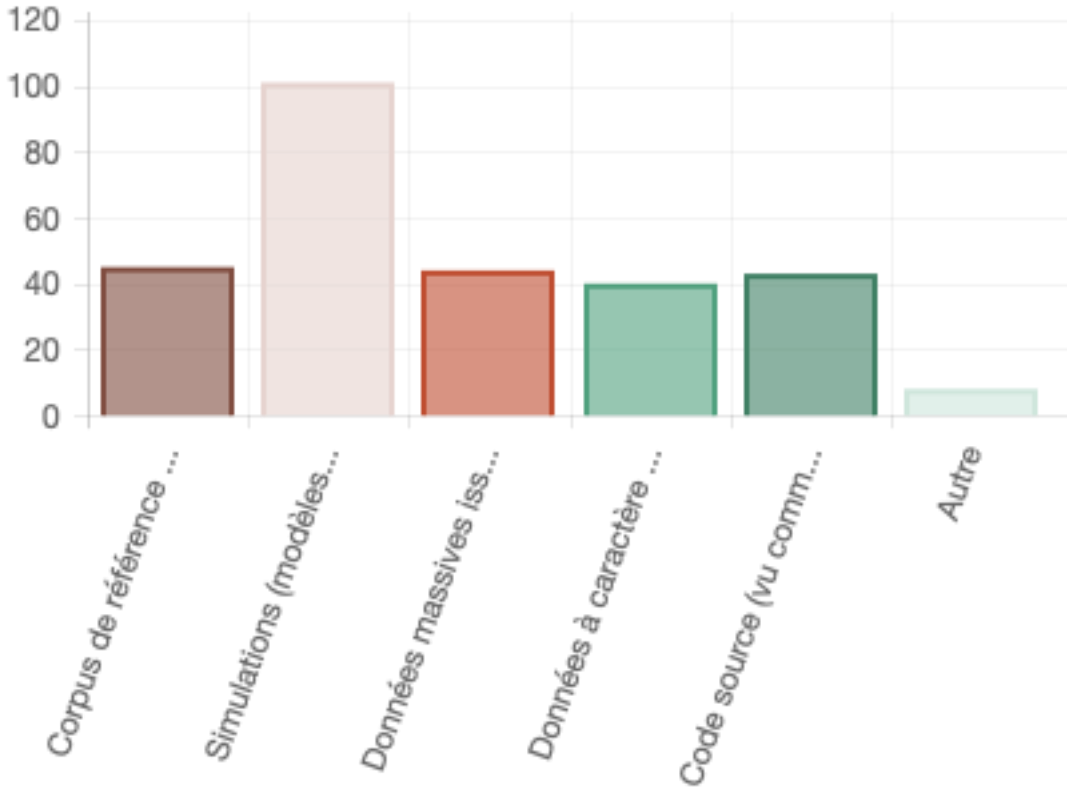
- 139 réponses envoyées, 22 réponses complètes non transmises, 7 réponses incomplètes
  - 122 réponses individuelles, 17 au titre de l'équipe
- Couverture scientifique: 115 équipes
  - Bordeaux: 15 équipes représentées
  - Grenoble: 21
  - Lille: 9
  - Nancy: 11
  - Paris: 14 (+2 Paris-Saclay)
  - Rennes: 9
  - Saclay 14 (+1 Saclay-Sophia)
  - Sophia: 17 (+1 hors équipe)
  - 1 hors équipe

Reflète la richesse des thématiques Inria

Algèbre linéaire numérique, Algorithmique arithmétique, Analyse de données, Analyse de données à grande échelle, Analyse de signaux EEG, MEG, Interfaces Cerveau Ordinateur, Analyse numérique, calcul scientifique, Modélisation, Simulation HPC, architecture des calculateurs, Assistance à la personne, Automatique, Bioinformatique, biomathématiques, microbiologie, Biologie computationnelle, biologie des systèmes, Biologie numérique, Biomécanique, Biomedical Engineering, Biostatistique, Calcul formel, Calcul intensif / HPC, Calcul parallèle, calcul scientifique HPC, Cancérologie, Compilation, Computational geometry and algebra, computer vision, Cybersécurité, data, mining, decentralised, communication networks, Distributed systems, Environnement, Évaluation et optimisation de performances de grandes infrastructures de calculs, fouille de données biomédicales, apprentissage, représentations des connaissances, Génie Logiciel, bases de données, Conception de langage, géométrie et topologie algorithmiques, Géométrie, Algèbre, Modélisation, Gestion et protection des données personnelles, Haptics, parallelisme, algorithmique, IHM , IHM visualisation, imagerie médicale, Intelligence Artificielle, optimisation, apprentissage, Intelligence artificielle, science des données, langages de programmation, les interfaces cerveau-ordinateur, Logique mathématique, Linguistique computationnelle, Machine Learning, mathématiques, mathématiques appliquées (simulation,), mathématiques appliquées et informatique, Mathématiques appliquées et simulation, électrophysiologie cardiaque, environnement, Mathématiques discrètes et codage, mathématiques, physique statistique, théorie des probabilités, statistiques, géométrie, anatomie computationnelle

Maths appliquées, Analyse et modélisation numérique, Mécanique des fluides, Propagation d'ondes,, Environnement, Calcul intensif et parallèle, Système d'information intégré, Micro-architecture, Modelisation neurosciences, Modélisation probabiliste, Modélisation stochastique, modelisation stochastique, apprentissage statistique/ profond, traitement d'image, teledetection,, imagerie de la peau, modelisation/optimisation, networks, Neuroinformatique, Neurosciences Computationnelles, Optimisation, Optimisation et complémentarité, ordonnancement pour le calcul parallèle, Perception interaction cognition, Preuves et vérification, Preuves formelles, Probabilités, Problèmes inverses, production de la parole, Réalité Virtuelle, Représentation des connaissances et raisonnement automatique, Réseaux, Réseaux informatiques, sécurité, Réseaux mobiles, Robotique , Robotique et intelligence artificielle (2), Santé, biologie et planète numériques / Sciences de la planète, de l'environnement et de l'énergie, Simulation numérique, Software Engineering / Programming Languages, Statistique, Neurosciences, synthèse d'images et acquisition numérique, systems, software engineering, TAL, théorie algorithmique des nombres, Théorie de jeux appliquée aux réseaux, Theory of control, Traitement automatique des langues, Traitement automatique du langage, traitement de la parole, traitement du signal, Traitement du signal audio, Traitement du signal et apprentissage, traitement du signal et machine learning, Véhicule autonome, Véhicule autonome et robotique mobile, Vérification et Preuves Formelles, Vérification formelle de protocoles cryptographiques, Vision artificielle, apprentissage automatique, Vision par ordinateur

Toutes types de données représentées



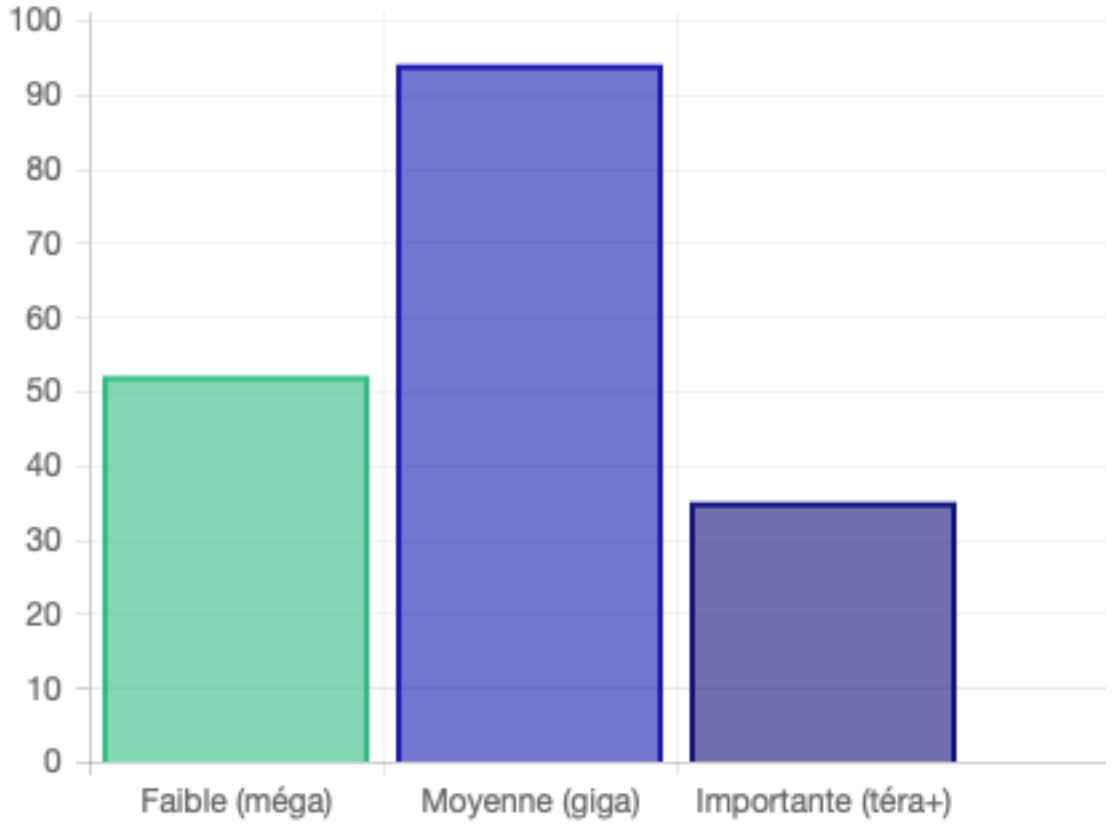
Pourquoi héberger?

Pourquoi avez-vous ou auriez vous besoin d'héberger ou de préserver vos données ?		
Comme base de vos travaux de recherche courants	115	82,73%
À des fins de reproductibilité	109	78,42%
Pour les réutiliser vous-même plus tard	105	75,54%
Pour accompagner vos publications	107	76,98%
Autre	5	3,60%

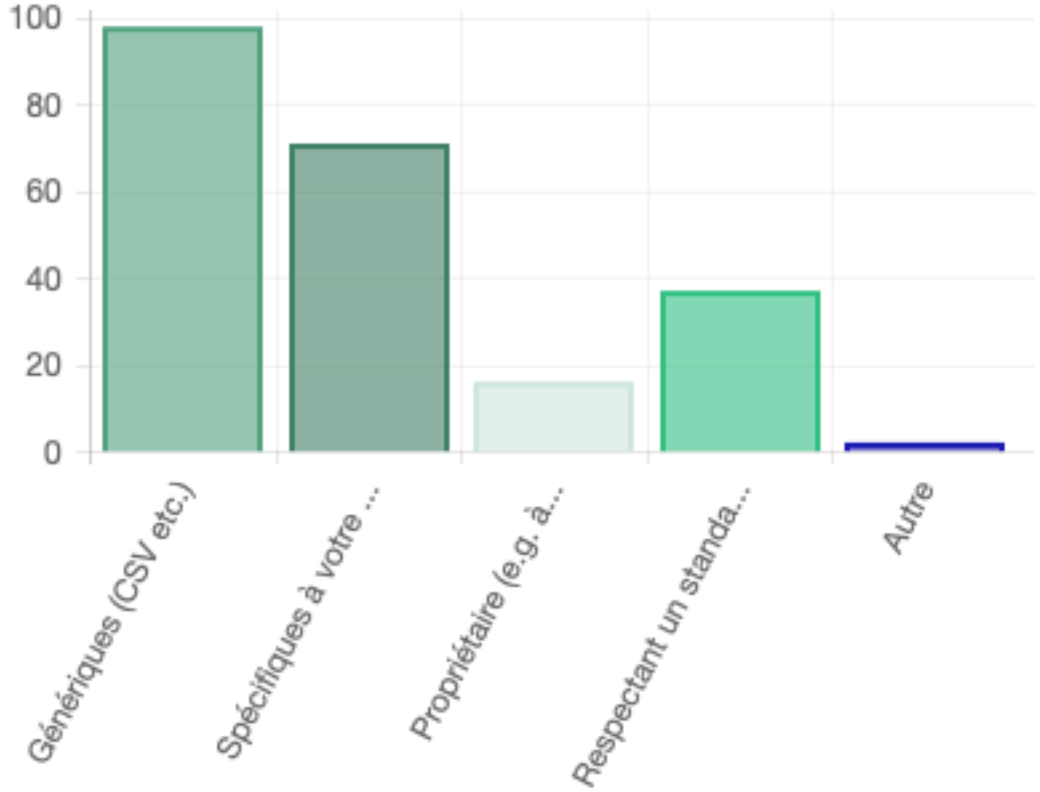
Pas d'idéologie particulière, mais un souci de garder une trace…

- Mention de la confidentialité
- Prise de conscience de la notion de justifiabilité (publications)

Taille des jeux de données



Formats utilisés



Cf. commentaires: très grosse variété des formats. Peu de signe d'une attention à la réutilisabilité (standards)

Documentation des jeux de données

Quelle documentation associez-vous à vos données ?		
Aucune (merci de préciser la raison)	20	14,39%
Spécifique à chaque jeu de données	82	58,99%
Indication du processus de création	52	37,41%
Sources et participants à la création du jeu de données	45	32,37%
Ajout de métadonnées intégrées aux données	41	29,50%
Rapport technique ou publication spécifique (e.g. data paper)	53	38,13%
Utilisation d'identifiants (identifiants thématiques, DOI, autres)	18	12,95%
Suivi à l'aide d'un carnet de laboratoire ou outil équivalent (Jupyter, ORG Mode sous Emacs etc.)	12	8,63%
Autre	4	2,88%

Plutôt rassurant

Faible

À améliorer…

Lien avec le code (commentaires dans le source), les fichiers paramètres, article spécifique

Hébergement

Comment hébergez-vous vos données ?		
Stockage de masse local (disque dur, CD etc.) (merci de préciser)	111	79,86%
Dans une plate-forme de partage: Git, github, gitlab (merci de préciser)	75	53,96%
Dans un cloud externe (merci de préciser)	13	9,35%
Dans une archive générique telle que Zenodo (merci de préciser)	12	8,63%
En accompagnement d'une publication déposée dans une archive ouverte telle que HAL (merci de préciser)	19	13,67%
En accompagnement d'une publication disponible sur le site d'un éditeur commercial (merci de préciser)	5	3,60%
Autre	9	6,47%

CD, DVDs, serveurs d'équipe
Gitlab et github sont +très+ utilisés
Présence croissante de Zenodo, référence à Huma-Num

Réutilisation et licences

Quelles conditions de réutilisation envisagez-vous ?		
Données largement ouvertes à tous	87	62,59%
Diffusion restreinte à une communauté scientifique	53	38,13%
Données diffusables sous condition car sensibles (droit d'auteur, données médicales, données personnelles, données protégées...)	34	24,46%
Données non diffusables	29	20,86%
Autre	4	2,88%

Utilisez-vous une licence particulière (e.g. Creative Commons) ?		
Réponse	Décompte	Pourcentage
oui (A1)	28	20,14%
non (A2)	111	79,86%

Contraste **ouverture** (engagée, cf. commentaires) – **licence** (CC-BY, logiciel)

En pratique

Science ouverte et accès ouvert

- Science ouverte *old school* : libre accès aux publications de recherche
  - Une étape indispensable avant de faire de la science ouverte en haute mer
  - La face immergée du travail scientifique, ce à quoi on va faire référence *in fine*
- Difficultés :
  - accès (paywalls)
  - coûts (abonnements ou APCs – revues hybrides)
  - réutilisation (e.g. fouille)
  - Souveraineté : disposer d'un corpus fiable, accessible et pérenne de ses production (chercheurs, communauté, institution, financeur)
- Exemple de politique d'établissement pour réfléchir: Inria
  - Obligation de dépôt dans HAL (lien avec les rapports annuels), même pour les articles dans des revues sous APC
  - Budget centralisé des APC – interdiction de l'hybride
    - OpenAPC
  - Désabonnements et réinvestissement dans le fonds pour la science ouverte

Science ouverte et gestion des données

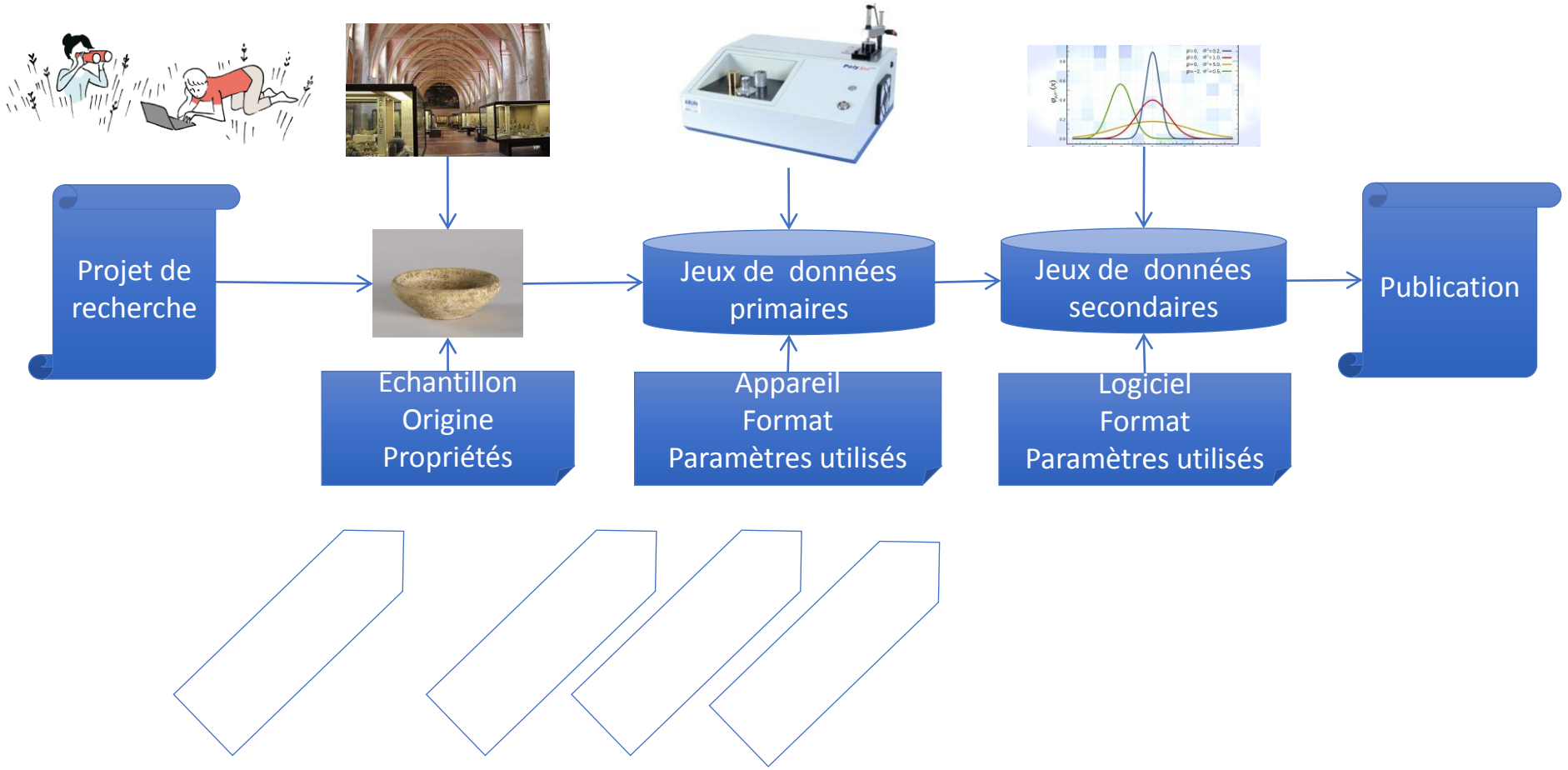
- Un passage obligé, le plan de gestion de données
  - Réfléchir en amont au conditions de production, de gestion, d'hébergement et de diffusion des jeux de données au cours du cycle de vie du projet recherche
  - Obligation ANR, Union Européenne, …
  - Un outil en ligne : DMP OPIDoR
- Quelles spécificité du domaine des matériaux anciens et patrimoine ?
  - Plongement dans le cycle de vie des données…

Eléments généraux d’un plan de gestion des données

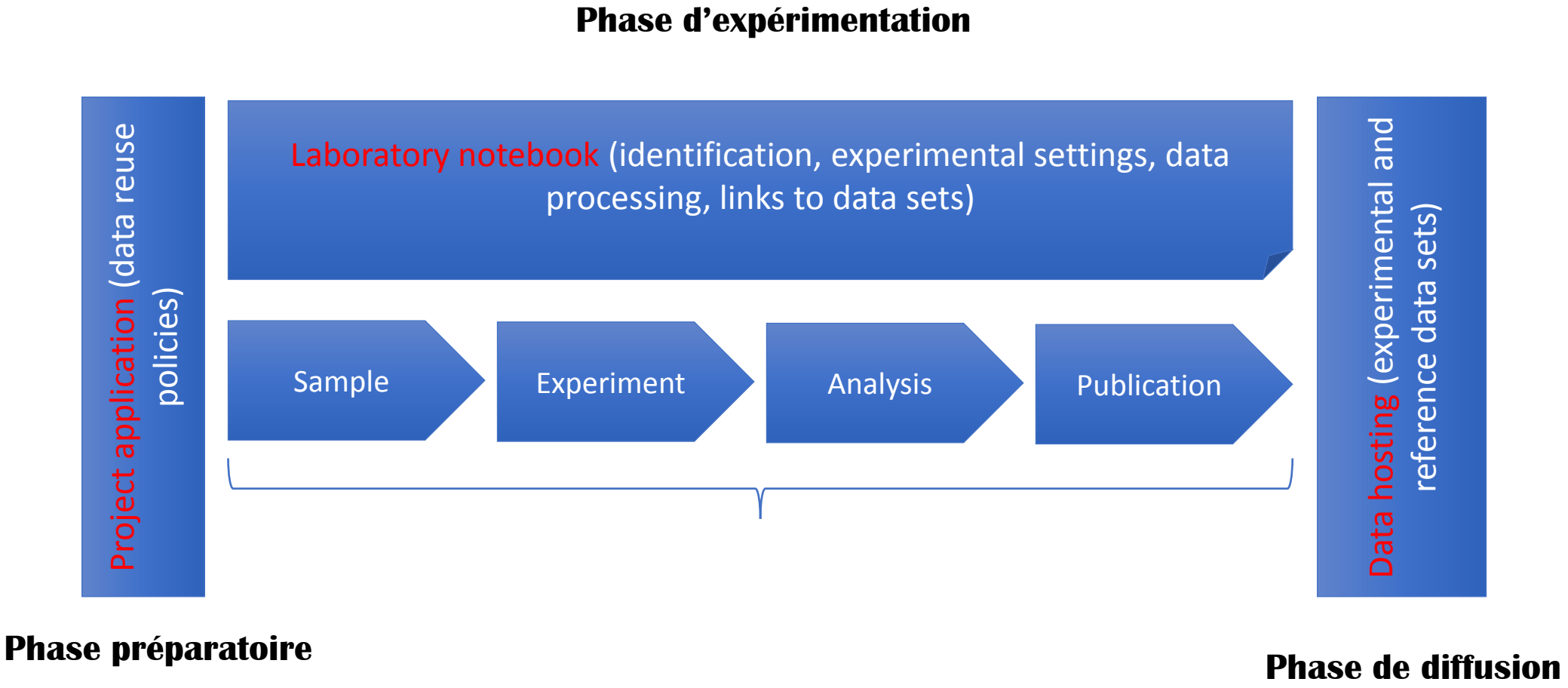
- Cf. modèle ANR
  - Une analyse à faire dans tous les cas – même si le DMP n’est pas imposé

=> expérience de Gwendoline Torterat

Tracer la création des jeux de données en sciences du patrimoine



Vers un environnement vertueux de gestion des données d'expérience



Documents et points de vigilance à chaque étape

- Phase préparatoire
  - Proposition de projet (financement, accès à un instrument)
  - Plan de gestion de données
  - Charte de réutilisation: accord amont entre les parties prenantes
- Phase d'expérimentation
  - Carnet de laboratoire pour documenter les différentes étapes de recueil de données et les éventuels calculs
    - Phase d'identification des acteurs intervenant dans le processus (PI, chercheur, technicien, stagiaire etc.)
    - Cf. MOOC science reproductible
    - Documentation globale des processus: http://ssk.huma-num.fr/#/
  - Hébergement des données temporaires, métadonnées associées, qualification pour une future sélection
- Phase de diffusion
  - Archivage et diffusion des documents intermédiaires
  - Sélection et stabilisation des données à vocation pérenne – détermination de leur degré d'ouverture

Un outil: la charte de réutilisation des données

- Contexte
  - Mise en place d'un mécanisme d'accord entre parties contribuant à la création d'un résultat numérique
    - Chercheur, organisation patrimonial, instrument, hébergeur de données
  - Travail initié en collaboration entre plusieurs grandes organisations Européennes
    - DARIAH, CLARIN, E-RIHS, APE, Europeana
  - https://datacharter.hypotheses.org
- Méthode
  - La charte est une déclaration faite par chacune des parties et affichée en tant que telle
    - Globale à une organisation, un projet, un échantillon ou un jeu de données particulier
- Principes
  - Reciprocity, Interoperability, Citability, Openness, Stewardship and Trustworthiness
    - https://datacharter.hypotheses.org/77
- Mise en œuvre:
  - Questionnaire destinés aux différents acteurs du processus de recherche
- Exemple
  - DARIAH-Campus

L’initiative DARIAH Campus

- Objectif: réunir en un même espace l’ensemble des ressources pédagogiques de DARIAH
- Une variété d’objet et de sources:
  - Vidéos, textes, reprises d’objets issus d’autres projets et par d’autres partenaires
  - Contenu en constante évolution
- Utilisation de la charte:
  - Définir des conditions générales de partage, réutilisation et contribution
  - Version complète sous:
    - https://campus.dariah.eu/docs/dariah-campus-reuse-charter
- Quelques éléments…

La charte de réutilisation de DARIAH campus

- Exposé introductif: donner une idée claire de la vision retenue
  - “Fostering **open access to scholarly resources**, as well as collaboration and fair data-sharing practices among a diverse range of actors involved in knowledge creation in the Arts and Humanities”
  - “as a **mutual declaration of goodwill**, the charter allows us to clarify our expectations regarding the interaction between content creators, users and curators”
- Un positionnement sur l’ensemble des 6 principes
- Une annexe spécifiant les détails techniques
  - Profil des métadonnées utilisées
  - Formats/standards utilisés pour les contenus (*en construction*)

Positionnement de DARIAH campus sur les 6 principes - 1

- Réciprocité
  - “learners ((re)users) are encouraged to share their **feedback** but also **contribute** new training materials”
  - “**contact information** of both authors (together with other contributors) and commenters/reviewers will be made explicit”
- Interopérabilité
  - “training materials will be made accessible in **open formats**”
  - “share metadata in a **standardized format** to ease harvesting our content”
  - “We make available the content of our hosted learning resources (written in Markdown, a lightweight markup language) available through a **GitHub repository**”
- Citabilité
  - [We] “make our recommended **citation model** explicit for each training resource shared via DARIAH-Campus”

Positionnement de DARIAH campus sur les 6 principes - 2

- Ouverture
  - “Open Educational Resource (OER)”
  - “openly and freely available under a **Creative Commons CC-BY 4.0 license**”
- Curation (*stewardship*)
  - “workflows to ensure the **long-term availability** of the training materials hosted on DARIAH-Campus”
- Fiabilité (*trustworthyness)*
  - “we recognize the diverse contributor roles (author, editor, contributor) to clearly document who participated in the production process in order to ensure its **traceability**”

Conclusions

- Différents outils à différentes étapes du projet
  - DMP: réflexion préalable au projet pour anticiper sur toutes les questions à se poser soi-même (donc, pas seulement une exigence administrative)
  - Charte: outil de dialogue entre partie et d'affichage d'une politique de réutilisation
  - Carnet de laboratoire: documentation fine des étapes menant à la production d'un jeu de données

Déposer ses publications dans HAL

Principes de base

- Une grande variété de documents
  - Article, poster, tutoriels, rapport, logiciel (en lien avec Software Heritage)
- A toutes les étapes
  - Manuscrit « diffusable », version soumise (preprint), version acceptée après révision (postprint), versions ultérieures corrigées
- Ajout de pièces complémentaires
  - Transparents de présentation à une conférence
  - Jeux de données
- Contenus modérés – une étape essentielle de la crédibilité de l'archive
  - Validation en surface du contenu
  - Equipe du CCSD, Inist et quelques établissements (e.g. AMU, Inria)
- Contenu organisé en *collections* et éventuellement accessible par le biais d'un *portail*

Les points d’entrée de HAL

- Différents portail, une seule base
  - HAL générique:
    - https://hal.archives-ouvertes.fr
  - Portails spécifiques
    - https://hal.inria.fr
    - https://halshs.archives-ouvertes.fr
    - …
- Un complément essentiel à HAL, AureHAL
  - Référentiels d’auteurs, de structures, de projets ANR, de projets Européens et de journaux scientifiques
  - https://aurehal.archives-ouvertes.fr

Se créer une identité numérique dans HAL

- Etape 1: se créer un compte
- Etape 2: se créer un IdHAL
  - Mon espace/mon profil
  - Pour les plus anciens:
    - Fusionner les vieilles identités sous HAL
    - Associer d'autres identifiants (VIAF, arXiv, ORCID etc.)
- Etape 3: se créer un CV (dès qu'on a une publication dans HAL)

Déposer un document dans HAL

- Déposer le document
- Compléter les métadonnées
- Points de vigilance:
  - « afficher la liste complète des métadonnées »
  - Licence
    - Privilégier CC-BY
  - Réutilisation des affiliations existantes (cf. AureHAL)
    - Eviter de créer des affiliations qui existent déjà